Published online 30 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W303–W307***
doi:10.1093/nar/gkn308

3D-Fun: predicting enzyme function from structure

Marcin von Grotthuss[1], Dariusz Plewczynski[2], Gert Vriend[1,*] and Leszek Rychlewski[2]

[1]CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Geert Grooteplein 26-28, 6525 GA Nijmegen, The Netherlands and [2]BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznan, Poland

Received February 29, 2008; Revised April 26, 2008; Accepted April 30, 2008

ABSTRACT

The 'omics' revolution is causing a flurry of data that all needs to be annotated for it to become useful. Sequences of proteins of unknown function can be annotated with a putative function by comparing them with proteins of known function. This form of annotation is typically performed with BLAST or similar software. Structural genomics is nowadays also bringing us three dimensional structures of proteins with unknown function. We present here software that can be used when sequence comparisons fail to determine the function of a protein with known structure but unknown function. The software, called 3D-Fun, is implemented as a server that runs at several European institutes and is freely available for everybody at all these sites. The 3D-Fun servers accept protein coordinates in the standard PDB format and compare them with all known protein structures by 3D structural superposition using the 3D-Hit software. If structural hits are found with proteins with known function, these are listed together with their function and some vital comparison statistics. This is conceptually very similar in 3D to what BLAST does in 1D. Additionally, the superposition results are displayed using interactive graphics facilities. Currently, the 3D-Fun system only predicts enzyme function but an expanded version with Gene Ontology predictions will be available soon. The server can be accessed at http://3dfun.bioinfo.pl/ or at http://3dfun.cmbi.ru.nl/.

INTRODUCTION

The advent of sequencing the human genome (1,2) has initiated a large number of genome-wide research efforts. Transcriptomics aims at generating the catalog of all expression patterns of genes as function of environmental parameters like cell-type, metabolic state, cell life cycle state, etc. Similarly, proteomics aims at this same catalog by directly measuring the proteins rather than indirectly via their expressed messenger RNAs. Metabolomics, glycomics, immunomics and pharmacogenomics, for example, aim at catalogs of all metabolites, all glycosylation sites and states, molecular functions associated with immune-related transcripts and a hosts' genetic response to drugs, respectively.

Structomics is the novel research field that aims at solving all protein 3D structures, and Table 1 shows a series of the better-known structural genomics efforts. As solving all ~25 000 human structures is not yet possible, efforts are directed towards first solving the 3D structures of enough significantly different protein sequences to be able to model the entire human protein structure repertoire by homology. This means that a series of protein structures is being solved for which the function cannot be estimated using a BLAST run against the databases of proteins with known function.

Many algorithms have been designed to aid with the detection of very remote sequence similarities. Examples are PSI-BLAST, several threading methods, profile–profile-based methods, etc. Structural genomics centers have deposited more than 1000 protein structures that have their function marked as 'unknown', and the function of more than 500 of those could not be determined using any of standard sequence-based methods. At http://kb.psi-structuralgenomics.org/KB/ a long list of such proteins is given.

It is commonly known that protein structure and the location of their functional sites are much more conserved than their sequences. Long after evolutionary divergence has made sequence similarity detection impossible, the structures of remote homologs are still very comparable. We here present a server, based on this concept, which helps with the detection of otherwise undetectable homologies. The 3D-Fun server takes as input the coordinates of a protein with known structure and unknown function, and compares it with all proteins with know structure available from the PDB. The output consists of a list of structurally similar proteins and with some vital superposition statistics. If the function of the database hits is known, it is listed to aid the user of the server with the prediction of the function of the query protein.

The 3D-Fun server is available at two sites: 3dfun.bioinfo.pl and 3dfun.cmbi.ru.nl, and a few more locations are planned. All underlying software is freely available too

*To whom correspondence should be addressed. Tel: +31 24 361 9521; Fax: +31 24 361 9395; Email: vriend@cmbi.ru.nl

Table 1. A few of the many structural genomics efforts

Consortium	www-page	No. of deposited PDP files
SGC	http://www.thesgc.com/	500
NESC	http://www.nesg.org/	500
TBSGC	http://www.doe-mbi.ucla.edu/TB/	500
SGXRC	http://www.nysgrc.org/	500
BSGC	http://www.strgen.org/	100
CESG	http://www.uwstructuralgenomics.org/	100
JCSG	http://www.jcsg.org/	600
MCSG	http://www.mcsg.anl.gov/	750
YSG	http://genomics.eu.org/spip/	25
RSGI	http://www.rsgi.riken.go.jp/	2000
SGPP	http://www.sgpp.org/	50
SECSG	http://www.secsg.org/	75
PSF	http://www.proteinstrukturfabrik.de/	20
SPINE	http://www.spineurope.org/	100

The three columns give the name of the consortium, their WWW-page and their stated number of deposited PDB files, respectively. Note that collaborations between centers in consortia may have caused double counting of PDB entries.

from the CMBI site. The 3D-Fun server has been set-up to predict the EC-code of an enzyme but it can, in principle, be extended to any other description of protein function.

METHODS

The 3D-Hit software (3) is used for all structure superpositions. This software superposes two structures at a time and determines a superposition score that is mainly based on the percentage residues that superpose within 3.0 Å (more detailed description of the 3D-Hit algorithm is presented below). An all-against-all structure comparison is performed and for each protein P with known function the superposition scores are determined for the most structurally similar protein Q1 that has a function different from P, and the worst superposing structure Q2 that superposes with a better score than Q1 and has the same function as P. This way the false-positive cutoff, being the average of the superposition scores for Q1 and Q2, can be determined for each protein in the database. This process is performed four times, once for each of the four functional levels of the EC-code. The final score for a comparison of an unknown protein with a protein in the database is determined relative to this false-positive cutoff value.

When querying with the protein R with known structure and unknown function the 3D-Hit program first compares R with a database of structures S_i that all have a known EC-code and can be characterized by four cutoff values. Sequence redundant proteins have been removed from this database. The EC-code of the protein S_i with the strongest structural similarity is assigned to the query if the superposition score is greater than all (or any) of the false-positive cutoff values. Let us, as an example, consider a query protein R which has a 3D-Hit score of 150 to protein S_i with EC-code 1.2.3.4, which has the four false-positive cutoff values 100, 120, 180 and 200. R will obtain an EC number assignment of 1.2.–.–.

All structural similarity scores are used for annotation in the 3D-Clust strategy (4). The query structure and all sequentially nonredundant proteins are hierarchically clustered by structural similarity using a complete link algorithm (5,6). The EC-code is completely (or partially) assigned to each group in each clustering iteration, if all of the enzymes in the group have the same function at all (or any) of the EC levels; otherwise, the EC-code is assigned as unknown. Let us, as an example, consider a cluster that contains four structures: the query protein and three enzymes with EC numbers 1.3.3.4, 1.3.3.6 and 1.3.4.1. This cluster will obtain an EC number assignment of 1.3.–.–. For the final prediction, the enzymatic function of the most detailed cluster that contains the query structure is used. Contrary to the 3D-Hit strategy, the 3D-Clust algorithm takes into account the enzymatic function of all structures S that have a better superposition scores with R than with all other proteins of the whole set.

The PD-Split software splits the PDB database in N equal chunks, where N is the number of available CPUs in the computer server that can be used by 3D-Hit. This way 3D-Fun can optimally use the power of coarse-grained parallel clusters.

The user needs to input the coordinates of the unknown protein in the well-known PDB format. Only the so-called ATOM records of the C-α atoms are required because the 3D-Hit (3) software used for the structural comparisons uses C-α coordinates only.

The server software sends the query structure to the N processors, waits for the results to come back, combines and sorts those results and presents them to the user. Figure 1 shows a typical server output page in which all important aspects of the output are annotated. On the Polish 48-node cluster a typical run takes 5–10 min.

The versions of 3D-Hit, PD-Split, that were optimized for the 3D-Fun server, and all necessary scripts to build your own 3D-Fun server are available from the CMBI site too.

3D-Hit algorithm

The structural similarity search algorithm, in general, is as follows. The query protein structure is dissected into overlapping fragments of 13 residues—equivalents of BLAST words. For each structural 'word', a set of template fragments is collected where (i) the residues in the centers are identical, where (ii) both distances between the first and the last C-α atoms are within 3 Å and where (iii) the RMSD after optimal superposition is below 3 Å. The rotation matrix and the translation vectors calculated during the superposition are used to rotate a 99-residues long segment of the query protein centered in the middle of the fragments. After the rotation, 99×99 dynamic programming matrix with spatial pairs of atoms denoted with 1 or 0 is created. A global alignment is conducted on the matrix. The aligned segments are superimposed and the calculated rotation matrix and translation vectors are used to compare fragments of 299 residues, in the same way as previously. The highest score obtained after the last alignment is used as a measure of similarity between the query and template proteins.

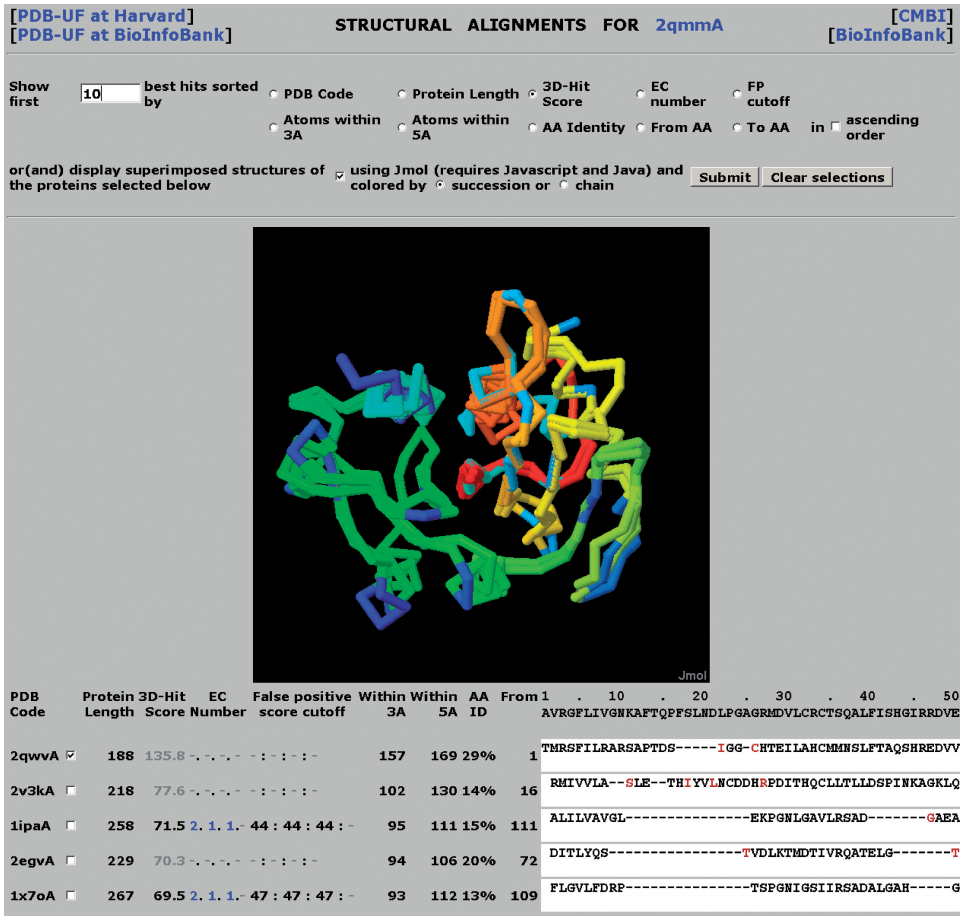

Figure 1. A screenshot of the 3D-Hit result list. In the top panel, the user selects the sort type and the maximum number of hits to be displayed. The backbone superposition of the query structure and the database hits selected by the user are displayed with Jmol (13). The residues are colored from N to C from red to blue. The database hits are shown below the Jmol viewer. If the superposition score is above the false-positive cutoff score, then the corresponding EC number is listed in blue color. Further details are explained at the server help-page.

Benchmark

The predictions generated by the 3D-Fun server are as good as structural alignments produced by the 3D-Hit program—the main core of the service. Here, we show the benchmark comparing results obtained using the 3D-Hit algorithm with results generated by the VAST program—one of the best-known programs for structural comparison of proteins. A data set of structural alignments between 45 537 protein chains generated by the VAST program was downloaded from the NCBI ftp server (ftp.ncbi.nih.gov/mmdb/vastdata). A total of 21 785 of those chains have annotated enzyme function in the corresponding PDB files. Protein sequences of these enzymes were clustered at the level of 90% amino acid identity using the CD-HIT sequence alignment program (7). Only one sequence from each cluster was used in the set of 3153 nonredundant proteins with known EC numbers and calculated structural VAST alignments. For each of those enzymes, we identified two proteins with the highest VAST alignment score; one of which has the same EC code and the other with a different one. The detection of the most structurally similar pairs of enzymes with the same and different function was then performed using the 3D-Hit program. We determined which alignment score best separates proteins with the same enzyme function from those with different EC codes. Figure 2 presents four ROC curves; each curve corresponds to one level of the enzyme classification. In this test, the 3D-Hit algorithm scored slightly better than the VAST program. Both methods outperformed results obtained with random alignment scores.

RESULTS

All structural genomics consortia together have deposited (28 February 2008) 6292 protein structures of which 1925 are labeled with 'Function unknown' and another 2411 have no EC-code given (provided numbers corresponds to sequentially redundant entries). Not having an EC-code obviously does not always mean that the function is not known, as many proteins simply do not have an EC-code associated or are not enzymes at all. Still, it was not difficult to make the short list of examples of proteins with know structure and unknown function that is shown in Table 2. In this table, we list five examples of proteins for which the 3D-Fun server can predict the function with decreasing degrees of certainty. The total CPU time spend on Table 2 was less than an hour.

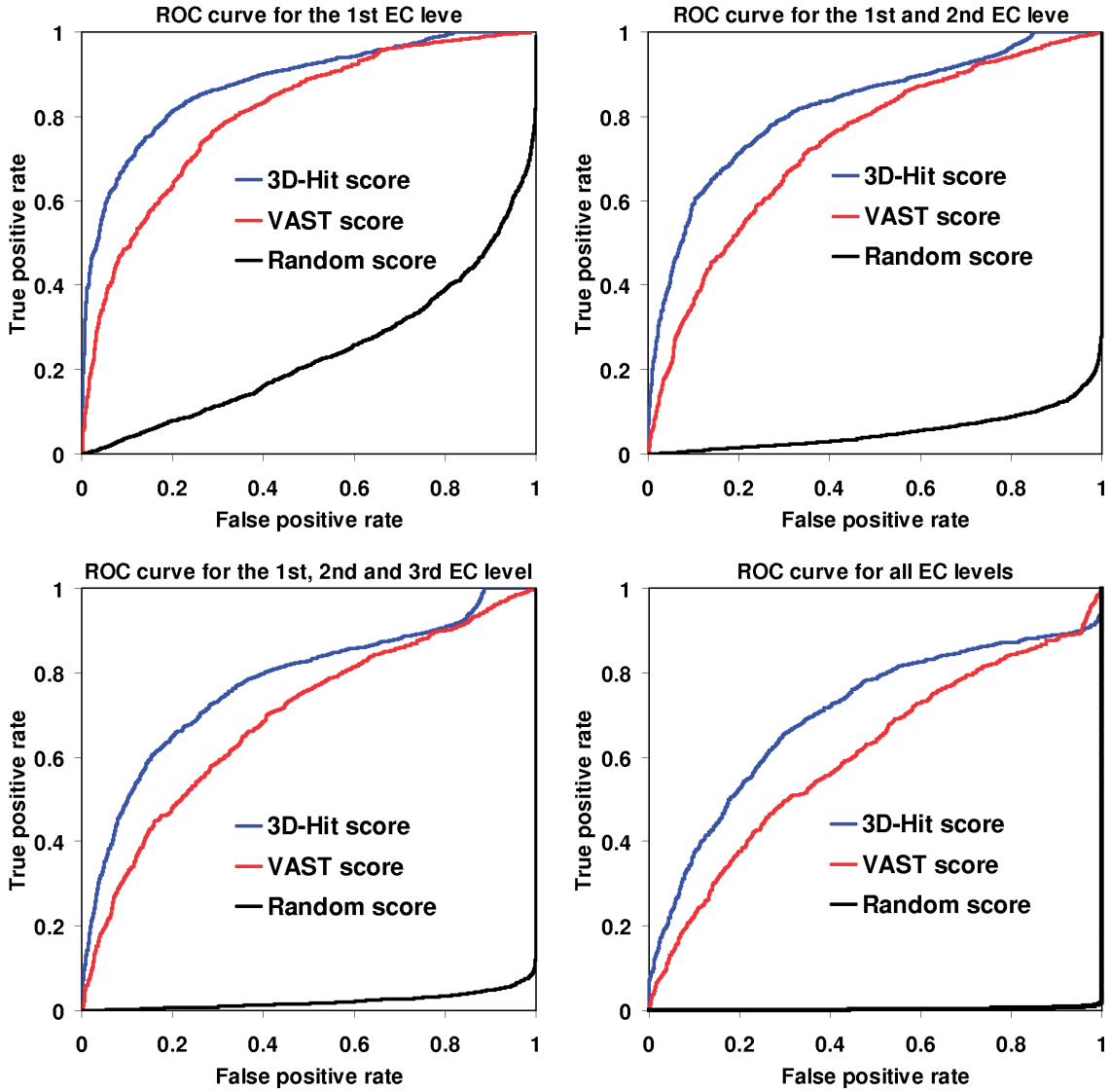

Figure 2. ROC curves for the 1st EC level (upper left chart); 1st and 2nd EC level (upper right chart); 1st, 2nd and 3rd EC level (lower left chart) and for all four EC levels (lower right chart). Note that the ROC curves for the random case (shown in black) are not diagonal lines as is usual in ROC plots. This is a consequence of the fact that prediction of enzyme function is a more difficult problem than bimodal classifications. Clearly, the probability of assigning an incorrect EC number in the random test is much bigger than assigning a correct one.

Table 2. Five examples of functional annotations made using 3D-Fun

No.	PDB accession codes	Predicted EC number	Predicted enzyme function
1.	2QMM, 2QWV	2.1.1.–	Methyltransferase
2.	1ZEE	1.3.11.–	Indoleamine 2,3-dioxygenase
3.	2G7Z	2.7.1.–	Phosphotransferase with an alcohol group as acceptor
4.	3BBJ	3.1. 2.–	Thioester hydrolase
5.	1YS9	3.1.3.–	Hydrolase

The predictions 1 and 2 are explained in greater detail in the text.

Example number 1 shows the prediction for two unpublished protein structures deposited by the MCSG (Table 1) that come from *Archaeoglobus fulgidus* and *Vibrio cholerae* species and have the PDB accession code: 2QMM and 2QWV, respectively. These two proteins belong to the PFAM (8) DUF358 family, which is a member of the α/β-knot clan. The DUF358 family contains a series of ~200 amino acids long archaeal and bacterial proteins. The function of these proteins is still unknown; however, they do contain several conserved histidine and aspartate residues that might form a metal-binding site. The 3D-Fun server predicted that these two proteins have a methyltransferase function (EC-code 2.1.1.–). Similar results were obtained in our previous studies, where we showed that four other hypothetical proteins (with the PDB accession codes: 1VH0, 1NS5, 1O6D and 1TO0) are probably also methyltransferase enzymes (4). Both predictions are supported by the fact that these six proteins share the same deep trefoil knot structure in the catalytic domain and have conserved a noncanonical AdoMet/AdoHcy-binding site (4).

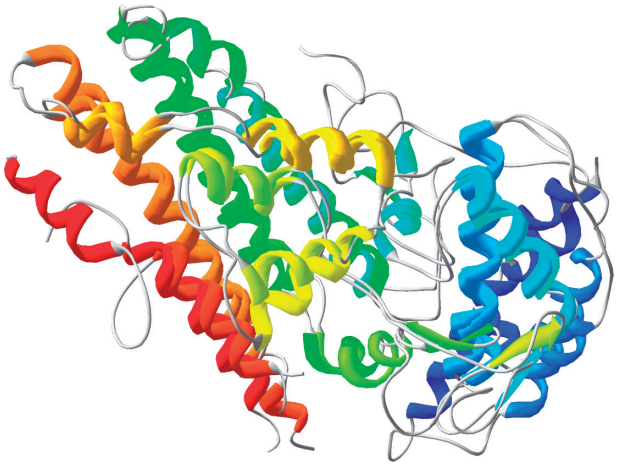

Figure 3. Superposition of a query structure 1ZEE and a protein structure of indoleamine 2,3-dioxygenase function with the 2D0T PDB code. These two proteins share only 17% sequence identity but over 66% of their C-α atoms are aligned and within 3 Å. The chains are colored from blue (N-termini) to red (C-termini).

Example number 2 presents the functional annotation for 1ZEE of the hypothetical protein SO4414 from *Shewanella oneidensis*, which has been deposited by the NSGC (Table 1). This protein is a member of the PFAM DUF1864 family, which still remains to be characterized. A 3D-Hit structural search detected a strong similarity to a indoleamine 2,3-dioxygenase family, represented by the 2D0T structure (9). The 3D-Clust program provided similar results by clustering the query model with three dioxygenase enzymes [2D0T, 1YW0 and 2NOX (10)] into one well-superposing group. Figure 3 presents the backbone superposition of the query structure and 2D0T.

DISCUSSION

Functional annotation using sequence similarity is very common practice. Indeed, the EXPROT project (11) showed that the function annotations in sequence databases are for the largest part obtained this way. The 3D-Fun server is heavily based on the concept that structure is more conserved than function, and that similar structures must indicate an evolutionary divergent relation and thus also functional similarity. Clearly, there are examples where these assumptions fail. The TIM-barrel motif, for example, has been associated with dozens of different functions (12). However, this is not as large as a problem as it initially looks because if structural comparison hits are found to proteins with different functions we know we cannot predict a function, and the server lists all hits with and concludes 'Function Unknown'. Things do go really wrong, however, when only one structure with known function is found in the database while multiple similar structures with very different functions are still waiting to be solved. Fortunately, such cases are most likely rare. Even if 3D-Fun predicts two different functions for a protein, the user is still helped because an experiment to validate a few function predictions is, of course, much simpler to perform than a totally *ab initio* experimental function determination.

At present it is still relatively easy to perform an all-against-all structure superposition for all proteins in the PDB. If, 1 day, the speed by which the structural genomics consortia solve their structures outruns the speed of our computers, then we can speed up the process by leaving out PDB files without known function or PDB files that are structurally and functionally similar to other PDB files already in the 3D-Fun database. Additionally, we can also try using the Grid rather than just one cluster computer. Efforts to run 3D-Fun on the Dutch national bioscience Grid are underway and results look promising.

ACKNOWLEDGEMENTS

M.v.G. acknowledges a fellowship from FEBS. G.V. acknowledges EMBRACE (LHSG-CT-2004-512092) and NBIC. D.P. acknowledges Diatomics (LSHG-CT-2004-512035) and MNiSW grant no. PBZ-MNiI-2/1/2005. L.R. acknowledges BioSapiens (LSHG-CT-2003-503265) and MicrobeArray (COOP-CT-2004-508399). Funding to pay the Open Access publication charges for this article was provided by BioSapiens.

Conflict of interest statement. None declared.

REFERENCES

1. McPherson,J.D., Marra,M., Hillier,L., Waterston,R.H., Chinwalla,A., Wallis,J., Sekhon,M., Wylie,K., Mardis,E.R., Wilson,R.K. *et al.* (2001) A physical map of the human genome. *Nature*, **409**, 934–941.
2. Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
3. Plewczynski,D., Pas,J., von Grotthuss,M. and Rychlewski,L. (2002) 3D-Hit: fast structural comparison of proteins. *Appl. Bioinform.*, **1**, 223–225.
4. von Grotthuss,M., Plewczynski,D., Ginalski,K., Rychlewski,L. and Shakhnovich,E.I. (2006) PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics. *BMC Bioinform.*, **7**, 53.
5. Defays,D. (1977) An efficient algorithm for a complete link method. *Comput. J.*, **20**, 364–366.
6. Murtagh,F. (1983) A survey of recent advances in hierarchical clustering algorithms. *Comput. J.*, **26**, 354–359.
7. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
8. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32(Database issue)**, D138–D141.
9. Sugimoto,H., Oda,S., Otsuki,T., Hino,T., Yoshida,T. and Shiro,Y. (2006) Crystal structure of human indoleamine 2,3-dioxygenase: catalytic mechanism of O2 incorporation by a heme-containing dioxygenase. *Proc. Natl Acad. Sci. USA*, **103**, 2611–2616.
10. Zhang,Y., Kang,S.A., Mukherjee,T., Bale,S., Crane,B.R., Begley,T.P. and Ealick,S.E. (2007) Crystal structure and mechanism of tryptophan 2,3-dioxygenase, a heme enzyme involved in tryptophan catabolism and in quinolinate biosynthesis. *Biochemistry*, **46**, 145–155.
11. Ursing,B.M., van Enckevort,F.H., Leunissen,J.A. and Siezen,R.J. (2002) EXProt: a database for proteins with an experimentally verified function. *Nucleic Acids Res.*, **30**, 50–51.
12. Watson,J.D., Laskowski,R.A. and Thornton,J.M. (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.*, **15**, 275–284.
13. Herráez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Edu.*, **34**, 255–261.